Journal of Biomedical Informatics 46 (2013) 294–303

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A semantic framework to protect the privacy of electronic health records with non-numerical attributes

Sergio Martínez, David Sánchez, Aida Valls *

Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain

ARTICLE INFO

Article history:
Received 22 June 2012
Accepted 18 November 2012
Available online 8 December 2012

Keywords:
Electronic health records
Privacy
Statistical disclosure control
Semantic similarity
SNOMED CT

ABSTRACT

Structured patient data like Electronic Health Records (EHRs) are a valuable source for clinical research. However, the sensitive nature of such information requires some anonymisation procedure to be applied before releasing the data to third parties. Several studies have shown that the removal of identifying attributes, like the Social Security Number, is not enough to obtain an anonymous data file, since unique combinations of other attributes as for example, rare diagnoses and personalised treatments, may lead to patient's identity disclosure. To tackle this problem, Statistical Disclosure Control (SDC) methods have been proposed to mask sensitive attributes while preserving, up to a certain degree, the utility of anonymised data. Most of these methods focus on continuous-scale numerical data. Considering that part of the clinical data found in EHRs is expressed with non-numerical attributes as for example, diagnoses, symptoms, procedures, etc., their application to EHRs produces far from optimal results. In this paper, we propose a general framework to enable the accurate application of SDC methods to non-numerical clinical data, with a focus on the preservation of *semantics*. To do so, we exploit structured medical knowledge bases like SNOMED CT to propose semantically-grounded operators to *compare*, *aggregate* and *sort* non-numerical terms. Our framework has been applied to several well-known SDC methods and evaluated using a real clinical dataset with non-numerical attributes. Results show that the exploitation of medical semantics produces anonymised datasets that better preserve the utility of EHRs.

1. Introduction

Recent advances in Electronic Health Record (EHR) technology have significantly increased the amount of clinical data electronically available. These data, consisting of medical and scientific documents and also of digitalised patient health records, are valuable resources for clinical and translational research. The analysis of the health care experience captured in clinical databases may lead to improved continuity in patient assessment, improved treatment, avoidance of adverse drugs reactions, and in ensuring that people at risk receive appropriate support services [11] [18].

Since medical information is usually associated to individuals, privacy must be ensured when data is made available for secondary use. This is explicitly stated by the Data Protection Act 1998 and the Human Rights Act 1998, which consider clinical data as "sensitive". Nowadays, EHRs are collected and maintained by public and private institutions that made them available for clinicians and researchers. Those institutions should guarantee that health information associated to patients is only made public with patient's authorisation. Exemptions are allowed in Section 39 of the Data Protection Act for medical purposes as well as statistical or historical research [11]. Moreover, the US Health Insurance Portability and Accountability Act (HIPAA) privacy rule permits publishing personal health information for public-health purposes without patient consent, if individual's privacy is "sufficiently" guaranteed [24]. To guarantee this privacy, the HIPAA either requires that experts in statistical and/or scientific methods certify that an individual's identity is protected from exposure under reasonable expectations (*Expert Determination*) or requires the removal of 18 data elements, named Protected Health Information (PHI) [1], to consider data de-identified (*Safe Harbour*); PHI includes names, census, geographical and financial information and biometrics, among others.

While data de-identification (by removing identifying attributes as proposed in the HIPAA privacy rules) prevents linking confidential data and patient's identity, it provides a false appearance of anonymity. Patient disclosure could still happen through statistical matching of remaining attributes. Several studies demonstrated that it is still possible to identify a patient by combining attributes which, when considered individually, did not seem problematic [5,11], that is, they were not considered PHI. There have been cases of disclosure in a priori protected clinical data, such as the

* Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Spain. Fax: +34 977 559710.

E-mail addresses: sergio.martinezl@urv.cat (S. Martínez), david.sanchez@urv.cat (D. Sánchez), aida.valls@urv.cat (A. Valls).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.11.005

identification of the clinician and the patient in a late abortion case through the analysis of released tabular data [31]. The identification was possible due to the low amount of late abortions in the region in which both the patient and the clinician were located. In [18] authors developed a technique for re-identifying seemingly anonymous genomic data by analysing combinations of, a priori, non-identifying attributes. Medical data, due to their variability and high dimensionality, is very prone to the appearance of identifying combinations of attribute values, which may enable disclosure when evaluated together.

The disclosure of patient electronic health records supposes a serious threat. Moreover, the awareness of the disclosure risks inherent to data release may lead to future reluctances and lack of trust in making data available for research. The fact that medical data does not get published or that it is excessively perturbed (e.g. encrypted or partially supressed) to minimise disclosure risk to the extent that it does not meet a level of accuracy, may produce a severe impact in its utility, hampering the benefits that can be extracted from its analysis [30].

To avoid these problems, privacy preserving methods providing additional privacy guarantees over the basic data de-identification must be developed. These methods suppose that direct or formal identifiers (such as PHI elements) have been already removed (e.g. according to the HIPAA privacy rules), and generate some distortion (i.e., masking) on the combinations of potentially identifying values. These last attributes are considered *quasi-identifiers*, referring to attribute sets whose values may produce rare or even unique combinations (e.g., very specific diagnostics or personalised treatments), and which may unequivocally identify an individual. Hence, these attribute can be used by a potential attacker to disclose the identity of certain individuals. In addition to data masking, it is also equally important to assure the quality of the published data. Therefore, data distortion derived from quasi-identifiers masking should be done in a way the anonymised data retains its utility as much as possible, so that similar conclusions can be extracted from the analysis of the original and the anonymised version of the dataset.

Different techniques for masking quasi-identifiers can be envisioned according to the type of data to deal with. Medical data can be presented as unstructured textual documents or as structured patient records collecting values for a set of normalised attributes (e.g., symptoms, diagnosis and treatment observed in a visit to a certain patient). In the former case, document sanitisation methods have been developed [24]; the latter case refers to privacy protection in structured databases [5], which is the focus of this paper.

1.1. Statistical disclosure control in structured databases

o minimise disclosure in structured databases (like EHRs), anonymisation methods framed in the Statistical Disclosure Control (SDC) area have been proposed. Many of these methods focus on removing/modifying quasi-identifiers, such as diagnosis + treatment, to fulfil a privacy model such as *k-anonymity* property [37], in which this paper is focused. An anonymised dataset is considered k-anonymous if any combination of quasi-identifier attribute values is repeated, at least, k times. The practical result is that each record (e.g. a patient health record) in a dataset is indistinguishable from, at least, $k-1$ other records with respect to their quasi-identifier attributes. The higher the k-value is, the lower the chance that an attacker could disclose individual identities, even when applying statistical inference over attribute value combinations.

Since original data have to be modified so that it becomes k-anonymous (and also more homogenous), a *loss of information* occurs. High information loss in anonymised datasets is undesirable, since it may negatively influence data utility in certain analytical tasks [5]. Hence, considering that different masking processes could equally achieve a desired level of k-anonymity, SDC methods select the one that, according to a heuristic, minimise the information loss. In general, information loss and disclosure risk are opposite dimensions that SDC methods try to optimise/balance [5].

Different techniques can be identified according to the algorithmic principles in which SDC methods rely to create anonymised datasets [5]. The simplest methods *supress* records whose attributes represent unique or rare value combinations [32,36]. This strategy has been usually applied in the past to anonymise medical data [11,28]. Even though, it results in a high information loss since the utility of supressed records is completely lost. Moreover, in heterogeneous datasets (such as health records), many records could be removed.

Instead of removing, more sophisticated methods modify attribute values to make records indistinguishable (i.e. k-anonymous) from other ones. The most well-known is *microaggregation* [2,6,8,19]. It builds groups/clusters of, at least, k original records according to a similarity function; then, each record of each cluster is replaced by the centre of the cluster, obtaining a k-anonymous dataset with the *same* number of records as the original one. Since similar records are aggregated together, the information loss resulting from the replacements can be minimised. Other methods are based on data *resampling* [7,13], which sample and sort the records (according to a comparison criterion), and replace each one by the average of the values of all taken samples. By taking and replacing, at least, k records/samples at each iteration, k-anonymity can be fulfilled [20]. Finally, *recoding* methods [21] iteratively replace the most similar record pairs/sets so that they become indistinguishable. The process is repeated until the whole dataset becomes k-anonymous.

Most algorithms focus on numerical attributes [5]. When applied to numerical data, the goal of anonymisation methods is to minimise the disclosure risk while retaining the distributional and statistical features of original data. In recent years, however, large amounts of non-numerical data, such as categorical attributes or textual responses, are commonly collected and published [39]. In the medical domain many potentially identifying data is expressed by means of non-numerical attributes, such as textual visit outcomes, diagnoses or treatments [24]. The accurate management (comparison/transformation) of this kind of data is not straightforward because, on the contrary to numbers, they take values from a discrete and finite list of modalities, which are usually expressed by words. Since arithmetic operators cannot be applied to this kind of data, simplistic approaches use equality/inequality operators and distributional statistics (e.g., mode) to compare and aggregate them [6,9,38]. These approaches neglect the most important dimension of non-numerical data, which is the way in which humans interpret them: *semantics*. Since the preservation of semantics is crucial to ensure the utility of anonymised results [22,39], these methods are likely to severely hamper the utility of anonymised data.

Recently, some authors have started considering data semantics during anonymisation [12,19,21,39]. Since semantics is an inherently human feature, the interpretation of textual values requires some sort of human-defined knowledge source, which provides a formal and machine-readable way to express a shared conceptualisation by means of a unified terminology and semantic interrelations (e.g., taxonomical generalisations). Ad-hoc structures or more general domain taxonomies, folksonomies, or ontologies [2,4,12,22,25] have been used to accurately anonymise textual data and maximise its utility by retaining data semantics.

Due to the importance of terminology and knowledge in clinical assessment, the medical domain has been very prone to the development of large and detailed knowledge structures. ICD-9/10, MeSH or SNOMED CT [26,35] are paradigmatic examples of

structured medical terminologies, containing thousands of taxonomically structured medical terms. However, as far as we know, there are no precedents using these structures to semantically anonymise non-numerical medical data.

1.2. Contribution and plan

In this paper, we propose a general framework that enables the anonymisation of structured non-numerical medical data (i.e., medical terms, such as names of symptoms or diagnosis) from a semantic perspective. The framework proposes and formalises three operators (*comparison*, *aggregation* and *sorting*) that, by exploiting medical knowledge structures (like the above-mentioned ones), enable a semantically-coherent managing of medical terms.

Afterwards, the framework is used to adapt three well-differenced SDC methods, so that structured non-numerical data could be *k*-anonymised while retaining their semantics as much as possible. These methods are evaluated and compared using a real medical dataset consisting of structured clinical outcomes and SNOMED CT as the medical knowledge base, under the dimensions of semantic information loss and computational cost. Results show that a semantically-grounded approach will more likely retain the utility of anonymised data when compared with classical non-semantic methods.

The rest of the paper is organised as follows. Section 2 presents our framework, formalising three semantically-grounded operators. Section 3 shows how this framework can be applied to several well-known SDC methods. Section 4 evaluates our proposal using those methods. The final section contains the conclusions.

2. A semantic framework for non-numerical data

As discussed in Section 1.2, classical SDC methods omit the semantic component of non-numerical data, focusing solely on their distributional properties. In this section, we present a general framework that integrates both features of non-numerical data. The framework captures the semantics of data by relying on the measurement of the *semantic similarity* between terms [33], assessed from medical knowledge bases, while taking into account the frequency of data, as well. In the following, we formalise the problem to solve; then, three semantically-grounded operators to manage non-numerical data are presented: *comparison*, *aggregation* and *sorting*.

2.1. Problem formalisation

A structured database consists of n records corresponding to individuals, each one containing m attribute values. To explicitly consider term frequencies, we represent the dataset in the form of $V = \{\langle v_1, \omega_1\rangle, \ldots, \langle v_p, \omega_p\rangle\}$, where each $\langle v_i, \omega_i\rangle$ tuple states the number ω_i of repetitions of each *distinct value* v_i found in V. Note that, typically, p (i.e., the number of distinct values) would be significantly lower than n (i.e., the total amount of records).

Example 1. Given the univariate dataset V stating patient diagnosis, {*asbestosis*, *degenerative disorder*, *amyotrophia*, *myofibrosis*, *asbestosis*, *allergy*, *myofibrosis*, *allergy*, *squint*, *amyotrophia*, *degenerative disorder*, *allergy*}, we represent it as $V = \{\langle asbestosis, 2\rangle, \langle degenerative\ disorder, 2\rangle, \langle amyotrophia, 2\rangle, \langle myofibrosis, 2\rangle, \langle allergy, 3\rangle, \langle squint, 1\rangle\}$. Note that, in this example, $m = 1$, $n = 12$ and $p = 6$.

This formalisation can be generalised for multivariate datasets with $m > 1$ attributes as follows. Let $MV = \{\langle\{v_{11}, \ldots, v_{1m}\}, \omega_1\rangle, \ldots, \langle\{v_{p1}, \ldots v_{pm}\}, \omega_p\rangle\}$ be the representation of the dataset, where each tuple $\{v_{i1}, \ldots, v_{im}\}$ represents a *distinct combination* of m attribute values, and ω_i states its number of occurrences (i.e., the frequency).

Considering that the goal of many SDC methods (like those introduced in Section 1.2) is to fulfil the *k*-anonymity property, if ω_i is equal or greater than a given value of k, the corresponding records in this tuple are already k-anonymous since they fulfil desired level of privacy. Hence, the goal of the anonymisation process consists of generating a dataset where $\omega_i \geqslant k$, $\forall i$.

2.2. Comparison operator

During the grouping step of the anonymisation process, similar records should be put together so that the information loss of the posterior value replacement could be minimised. To do so, a *comparison* operator is needed. Groups are built around a base value b, whose selection depends on the anonymisation method (more details in Section 3), to which the most similar records of the dataset are joined. Hence, the comparison operator should be able to rank the set of records in the dataset according to their *distance* with the base value b.

To consider the value semantics during this comparison, we rely on the notion of *semantic similarity*, which quantifies the taxonomical resemblance of compared terms based on semantic evidences extracted from a knowledge base [3,29,33]. To evaluate clinical terms, we use a medical structured terminology like SNOMED CT as the knowledge base, which offers a taxonomic structure in which subsumption relations are modelled as links between clinical terms. In [3,34], a state-of-the-art measure based term subsumption is proposed, which quantifies the *semantic distance* between term pairs $sd(v_1, v_2)$ as a function of the number of non-common subsumers of (v_1, v_2) (numerator in the fraction of Eq. (1)). This value is normalised by the complete set of subsumers of (v_1, v_2) (denominator in the fraction of Eq. (1)).

$$sd(v_1, v_2) = \log_2\left(1 + \frac{|T(v_1) \cup T(v_2)| - |T(v_1) \cap T(v_2)|}{|T(v_1) \cup T(v_2)|}\right), \quad (1)$$

where $T(v_i)$ is the set of taxonomic subsumers of v_i, including itself.

An advantage of this measure is that it evaluates *all* the taxonomical ancestors of the evaluated terms, considering also multiple taxonomical inheritance, which are very common in medical taxonomies [3]. As a result, it showed improved accuracy over related works in several medical [3] and general purpose [34] benchmarks.

Example 2. As an illustrative example of the semantic distance measure, let us consider a univariate dataset where the attribute refers to diseases: V_1 = {*asbestosis*, *amyotrophia*, *myofibrosis*, *allergy*, *degenerative disorder*, *squint*}. Fig. 1 shows an extract of the taxonomy modelling these diseases in SNOMED CT. Applying Eq. (1) to all the possible pairs of terms we obtain the semantic distance values shown in Table 1. We can see, for example, that

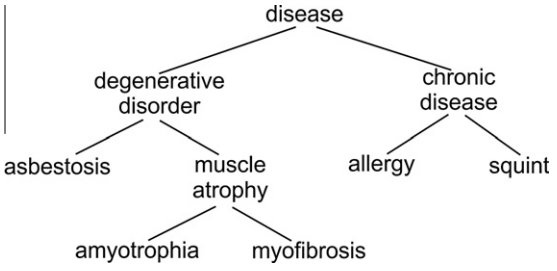

Fig. 1. The subsumer hierarchy for the set V_1, extracted from SNOMED CT.

Table 1
Semantic distance between term pairs of Example 2, according to the SNOMED CT taxonomy extract shown in Fig. 1.

Semantic distance	Asbestosis	Amyotrophia	Myofibrosis	Allergy	Squint	Degenerative disorder
Asbestosis	0	0.68	0.68	0.85	0.85	0.42
Amyotrophia	0.68	0	0.49	0.87	0.87	0.58
Myofibrosis	0.68	0.49	0	0.87	0.87	0.58
Allergy	0.85	0.87	0.87	0	0.58	0.81
Squint	0.85	0.87	0.87	0.58	0	0.81
Degenerative disorder	0.42	0.58	0.58	0.81	0.81	0

"sibling" terms like *amyotrophia* and *myofibrosis* are less distant than *allergy* and *squint*, because the former are more specific and, hence, they share more subsumers.

Applying this measure, we can semantically compare, rank and group the most similar (i.e., least distant) record values v_i in a dataset with respect to a base value b from which each group is built, using $sd(b, v_i)$.

To consider also the data distribution, since each distinct value v_i appears ω_i times in the dataset (as formalised in Section 2.1), we propose to count the semantic distance between a given value v_i and the base value b as many times as indicated by its frequency of appearance ω_i. In that way, the accumulated distances resulting from grouping together b and all the records with the value v_i can be minimised.

Formally, the *comparison operator* used to group records with respect to a base value is defined as follows.

Definition 1. The weighted semantic distance (*wsd*) between a univariate reference value b and a univariate set of records $\langle v_i, \omega_i \rangle$ is defined as:

$$wsd(b, \langle v_i, \omega_i \rangle) = \omega \cdot sd(b, v_i) \tag{2}$$

This measure can be generalised to multivariate data as follows:

Definition 2. The distance between a multivariate reference value with m attributes $\{b_1, \ldots, b_m\}$ and a multivariate set of records $\langle \{v_{i1}, \ldots, v_{im}\}, \omega_i \rangle$ is defined as the *average* of the weighted semantic distances of the individual attribute values:

$$wsd(\{b_1, \ldots, b_m\}, \langle \{v_{i1}, \ldots, v_{im}\}, \omega_i \rangle) = \frac{\sum_{j=1}^{m} wsd(b_j, \langle v_{ij}, \omega_i \rangle)}{m} \tag{3}$$

2.3. Aggregation operator

Aggregation refers to the process of replacing several records by a single one that summarised them, so that values becoming indistinguishable. The aggregated value is understood as the prototype or *centroid* of a set of values. Since the replacement of record sets by their centroids causes a loss of information, the selection of an accurate centroid is crucial to retain the utility of anonymised data.

Centroid calculus for numerical data relies on standard averaging operators (e.g. arithmetic mean) [5]. However, the accurate centroid calculus for non-numerical data is challenging due to the lack of semantic aggregation operators and the necessity of considering a discrete set of centroid values. Related works propose methods to compute centroids for non-numerical data *either* relying on the distributional features of data, where the centroid is the *modal* value [38], *or* on background semantic, where the centroid is the term that generalises all aggregated values in a taxonomy [2]. Since only one dimension of data (distribution or semantics) is considered, both approaches result in suboptimal results [23].

In this section, we propose a centroid calculation method for multivariate non-numerical data that considers, in an integrated manner, *both* semantics and distribution of data. To obtain accurate centroids, the background knowledge base is exploited not only to semantically compare terms, as proposed in the previous section, but also to retrieve the centroid candidates.

Formally, given a distance function d, the centroid of a set of values $\{v_1, v_2, \ldots, v_p\}$ is defined as:

$$centroid(v_1, v_2, \ldots, v_p) = \arg\min_{c} \left\{ \sum_{i=1}^{p} d(c, v_i) \right\} \tag{4}$$

where c is a centroid candidate for the set of elements.

The first issue concerns the search space of centroid candidates (c). Since c must be necessarily a discrete value, some approaches (like the ones based on taking the modal value of a sample [38]) bound the set of possible candidates to those values appearing in the input dataset. Hence, the centroid accuracy would depend on the granularity and suitability of the input data. In our case, since we rely on medical knowledge bases like SNOMED CT, which offer detailed and fine grained taxonomical structures, we extend the centroid search space to all terms of the taxonomy related to the input data. For semantic coherence, centroid candidates will be all input terms together with their taxonomical ancestors.

To apply Eq. (4) to the centroid calculation, it is necessary to use an appropriate distance function. We propose the use of the weighted semantic distance defined in the previous section.

Formally, let us take $V = \{\langle v_1, \omega_1 \rangle, \ldots, \langle v_p, \omega_p \rangle\}$ as an input dataset with a single non-numerical attribute. The first step maps the terms v_i of the set V in a background knowledge base (like SNOMED CT) and extracts the minimum hierarchy H that taxonomically models all v_i values. All terms in H, which include both values in V and their taxonomical ancestors, are considered as centroid candidates. Next, applying Eq. (4), the term c in H that minimises the weighted semantic distance (Eq. (2)) to all v_i in V will be selected as the centroid. Note that, in this case, each centroid candidate c acts as the base value in Eq. (2).

Definition 3. The centroid of a set of non-numerical values v_i in V is defined as the term c_j that minimises the weighted semantic distance *wsd* with respect to all the values v_i in the space V.

$$centroid(V) = \left\{ \arg\min_{\forall c_j \in H} \left(\sum_{i=1}^{p} wsd(c_j, \langle v_i, \omega_i \rangle) \right) \right\} \tag{5}$$

Example 3. As an illustrative example, let us consider the univariate dataset: $V = \{\langle asbestosis, 2\rangle, \langle degenerative\ disorder, 2\rangle, \langle amyotrophia, 2\rangle, \langle myofibrosis, 2\rangle, \langle allergy, 3\rangle, \langle squint, 1\rangle\}$. By mapping these clinical terms in SNOMED CT, we are able to extract the minimum hierarchy H, shown in Fig. 2. Hence, the centroid candidates are those values in H. For each one, we compute the accumulated sum of weighted semantic distances over v_i, as stated in Eq. (5) (see numbers in brackets in Fig. 2). In this case, the value that minimises the distance against all vi is "*degenerative disorder*", since its accumulated sum of *wsd* is 6.81. This value reflects the *information loss* resulting from replacing values in V by their aggregation, which should be ideally minimised. So, applying Definition 3, *centroid*(V) = *degenerative disorder*.

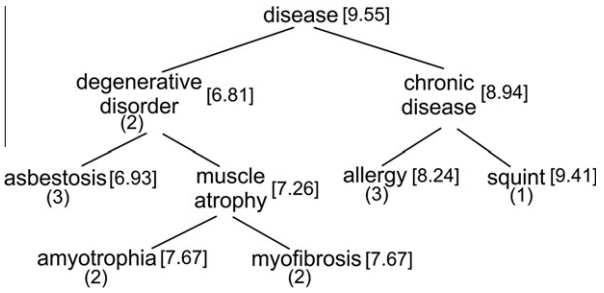

Fig. 2. The minimum hierarchy H for values in V, extracted from SNOMED CT. Numbers in parenthesis represent the amount of repetitions of each value in the dataset. Numbers in brackets represent the accumulated distance of each centroid candidate according to Eq. (5).

Notice that, in this example, neither the term that generalises all values in V (i.e. *disease*) nor the modal value (i.e. *allergy* and *asbestosis* with three appearances) minimise the accumulated distance. On the one hand, using as centroid the term that generalises all values for non-uniformly distributed data [2], usually results in a high loss of information due to the necessity to generalise outliers (e.g. *squint*). On the other hand, even though any of the modal values resulted in more accurate (but not optimal) centroids, the final selection will produce significantly different information loss (i.e. *allergy* with 8.24 and *asbestosis* with 6.93). The fact that several modal values exist is quite common in large datasets with a limited amount of modalities, a problem that affects approaches based solely on distributional features of data [38].

For simplicity, considering attributes as independent variables, the above method can be generalised for multivariate data by computing individual centroids for each attribute as follows:

Definition 4. The centroid of a set of multivariate data MV with m attributes is defined as:

$$centroid(MV) = \{centroid(A_1), centroid(A_2), \ldots, centroid(A_m)\} \quad (6)$$

where A_j the set the set of distinct values for the jth attribute in MV.

2.4. Sorting operator

Some SDC methods (like *resampling* [20]) require from sorting input data before grouping them. Again, sorting non-numerical values is not straightforward since, in general, they are not ordinal.

To sort a set of values, a reference point is needed, so that values could be arranged according to their similarity/distance to that reference. Numerically, this is done according to the max/min value (i.e. the most extreme value) of the set. To define a sorting procedure for the non-numerical case, we also rely on the notion of the most extreme value, which corresponds to the one that, globally, is the *most distant* to all other values (conceptually, this is the opposite of the centroid as computed in section 2.3). Once this reference value is obtained, other values are sorted by iteratively picking those that are least distant to that extreme value.

To obtain the reference value/record as well as to compare it to other elements in the set, we rely on the weighted semantic distance and the centroid calculus procedure proposed in Sections 2.2 and 2.3. The sorting procedure is formalised as follows:

Algorithm 1 (***Sorting procedure***).

1. Pick, as reference point, the **most distant** record (according to Eq. (2)) to the **centroid** of the dataset (obtained with Eq. (5)).
2. Sort the rest of records by taking, at each iteration, the **least distant** one (Eq. (2)) to the reference point.

Notice that Eqs. (3) and (6) should be used, instead, in case of multivariate datasets.

3. Applying the semantic framework to SDC methods

By means of the framework presented in the previous section and exploiting a medical knowledge base like SNOMED CT, non-numerical medical terms can be coherently compared and aggregated. Since many SDC methods rely on these operators to anonymise data, our framework enables an accurate anonymisation of medical datasets according to their semantic and distributional features.

In this section, we use our framework to adapt three well-known SDC methods (recoding, microaggregation and resampling, introduced in Section 1.2) originally designed to deal with numerical data, to the non-numerical case.

3.1. Recoding

Data *recoding* provides a simple way to build k-anonymous datasets. This methodology consists in replacing non-k-anonymous records (i.e., those whose $\omega_i < k$) by another already existing record with different attribute values [21]. To select the replacement that minimises the information loss, the *least distant* and *least frequent* distinct record pair is selected.

The behaviour for the recoding method presented in [21] is given in Algorithm 2, highlighting in **bold** the steps in which the operators proposed in our framework are used.

Algorithm 2 (***Recoding***).

1. Select the group of records with the minimum number of repetitions. As long as this number is lower than k, the dataset is not k-anonymous.
2. Find the **least distant** record to the records in the group, with the lowest amount of repetitions.
3. Original values are replaced by the **least distant** one, increasing their anonymity level due to the higher amount of value repetitions.
4. Repeat the process for the next least frequent record set (step 1) until the k-anonymity is fulfilled.

The adaptation of this method to non-numerical data is straightforward using the comparison operator proposed in the semantic framework in Section 2.2.

3.2. Microaggregation

Microaggregation is one of the most commonly used privacy-preserving methods, since it tends to better preserve data utility [2,6,16,19].

As microaggregation algorithm, we selected the well-known *Maximum Distance Average Vector* (MDAV) method [19]. It is based on generating clusters of at least k elements around the most distant records to the dataset centroid. To minimise information loss, clusters are built by picking up the least distant records around reference records. The method's behaviour is detailed in Algorithm 3, highlighting in **bold** the steps in which the operators proposed in our framework are used.

Algorithm 3 (***Microaggregation***).

1. Calculate the **centroid** of the complete dataset. Then, the **most distant** record r to the centroid and the **most distant** record s to r are selected.

2. A cluster is constructed by grouping, at least, the $k-1$ **least distant** records to r, obtaining a cluster **centroid**. The same procedure is repeated for the record s.
3. The whole process is repeated until no records remain ungrouped. Note than since it is possible that less than k records remain after clustering the dataset (i.e. they cannot form a k-anonymous cluster), these are added to the cluster with the **least distant** centroid, recalculating the **centroid** of the modified clusters.
4. Finally, original records are replaced by the centroid of the cluster to which they belong. Since all records are grouped and aggregated with at least $k-1$ other records, the resulting dataset is k-anonymous.

By means of our framework, whenever a centroid needs to be obtained (either of the whole dataset, step 1, or of a particular cluster, steps 2 and 3), the procedure proposed in Section 2.3 can be applied. Since clusters are built according to the selected centroid and records are replaced by cluster centroids, the fact that centroids minimise the accumulated semantic distances of the aggregated terms helps to minimise the information loss. Again, the comparison operator presented in Section 2.2, can be used to obtain the most/least semantically distant record, when needed. Since our operator considers all record value repetitions at once, identical records can be clustered together, obtaining more cohesive groups.

3.3. Resampling

Data *resampling*, as defined in [20], is an anonymisation method consisting in making k random samples of input data; then samples are sorted and records of each sample are grouped and aggregated so that they fulfil the k-anonymity property. Compared to the above methods, resampling is faster since the sampling process is done randomly, which makes it especially suitable for very large datasets. By contrast, the random criteria may negatively influence the information loss.

The method's behaviour is detailed in Algorithm 4, highlighting in **bold** the steps in which the operators proposed in the semantic framework are used.

Algorithm 4 (***Resampling***).

1. Create random k samples of n/k records, without replacement (i.e., each record is taken only once).
2. **Sort** these samples and create sets P_i with the records at the ith position of all samples, so that similar records are put together at similar positions of different samples.
3. Calculate the **centroid** of each P_i. Add the (n mod k) remaining records to the set P_i with the **least distant** centroid. The anonymised dataset is obtained by replacing all records of each P_i by the **centroid** of P_i. Since, by definition, P_i contains at least k records, this process generates a k-anonymous dataset.

In this case, the sorting procedure proposed in Section 2.4 can be applied to step 2 so that records are arranged according to their semantic similarity. The reference value to perform the sorting process is the centroid of each sample. Again, the proposed centroid procedure can be applied to compute semantically coherent centroids of each set (step 3), and the comparison operator can be used to select the least semantically distant centroid for the remaining records (step 3).

4. Evaluation

In this section, we evaluate the contribution of our framework to the SDC methods discussed in Section 3 during the anonymisation of clinical data with non-numerical attributes. Results are compared according to information loss against the classic non-semantic anonymisation and a naïve method based on data suppression.

Table 2
Example of clinical data used for evaluation. Numbers in parenthesis represent the ICD-9 codes.

ID	Age range	Patient ZIP code	Principal diagnosis cause of admission	Other condition that coexist at the time of admission
*	50–54	916**	Abstinent alcoholic (291.81)	Metabolic acidosis due to salicylate (276.2)
*	65–69	913**	Infected spinal fixation device (996.67)	Uric acid renal calculus (592.0)
*	65–69	903**	Aneurysm of thoracic aorta (441.2)	Cardiac oedema (428.0)
*	⩾85	902**	Fibroma of ovary (218.9)	Chronic osteoarthritis (715.9)
*	30–34	917**	Acute fulminating appendicitis (540.9)	Body mass index 40+ – severely obese (V85.4)

4.1. The dataset

As evaluation data, we used a structured database containing inpatient information provided by the California Office of Statewide Health Planning and Development (OSHPD) collected from licensed hospitals in California.[1] Specifically, we used the latest patient discharge dataset (4th quarter of 2009) of the hospital with the largest amount of records (i.e., Cedars Sinai Medical Center, Los Angeles County).

Prior to publication, the OSHPD has masked or removed some attribute values related to demographic variables that may enable re-identification (see an example on the first three columns of Table 2), as suggested by the HIPAA rules for data anonymisation [1]. For example, specific patient age has been masked in a range of 20 categories. Other attributes, such as the hospital identification number have been directly removed from the dataset, whereas the later digits of the ZIP code were also removed. However, clinical data related for example to diagnoses are published "as is", since these are not considered PHI; see the last two columns of Table 2. From these, we focused on two non-numerical attributes corresponding to the *principal diagnosis* and *other conditions* of the patient at the time of the admission (i.e., $m = 2$), which are stored as ICD-9 codes in the original data file. After removing records with missing information, a total of 3006 individual records is available for testing (i.e., $n = 3006$). Data distribution for these two attributes is shown in Fig. 3. A total of 2331 different combinations of values (i.e., $p = 2331$ tuples) can be found, from which a significant amount (2073) are unique. As demonstrated in previous works (such as the late abortion identification case [31]), when rare or even unique combinations of this type of clinical attributes appear, patient private information disclosure may happen if a third party knows other patient's data as for example, the hospital name, its address or the period of hospitalisation. Hence, we considered them as *quasi-identifiers* that should be masked in addition to the already removed identifiers suggested by the HIPAA privacy rules. At the same time, considering that patient diagnoses are valuable information for clinical research, its anonymisation should preserve the utility of data by minimising the information loss of the anonymisation process. Finally, since values for these attributes are non-numerical, a structured medical knowledge base

[1] http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html.

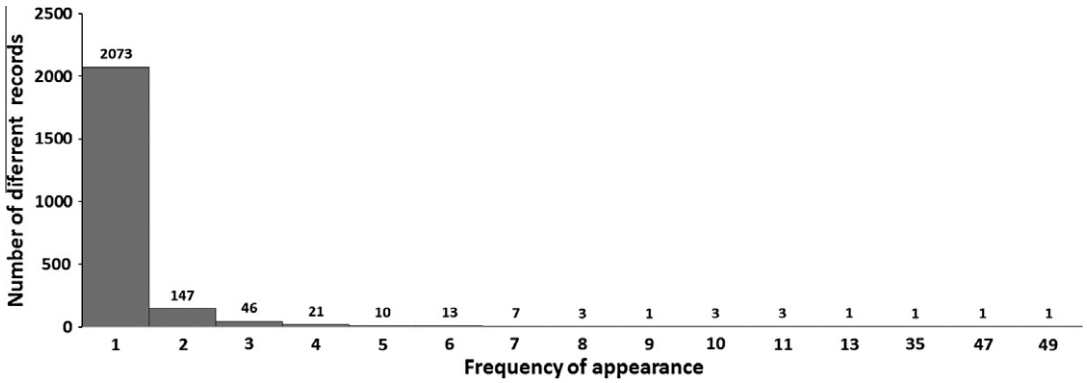

Fig. 3. Distribution of distinct value tuples for the *principal* and other *conditions* attributes.

has been used to extract and interpret their semantics, as proposed by our framework. Since ICD-9 codes were available for each condition, we translated them into SNOMED-CT concepts, using a publicly available mapping file.[2] After this mapping, the SNOMED-CT ontology can be used in the anonymisation and evaluation processes. Thanks to its size and taxonomic detail (with more than 311,000 unique concepts organised in 18 overlapping hierarchies with more than 1.36 million relationships), SNOMED-CT is especially suitable to assist semantic similarity assessments [3,29,33].

4.2. Evaluating the preservation of data semantics

Taking into consideration that the preservation of data semantics is crucial to retain the utility of anonymised non-numerical data [22,39], in our tests, we measured the quality of anonymised data by quantifying the *semantic information loss* caused by replacing quasi-identifier attribute values by their masked versions. Note that, since our point of departure is an already de-identified dataset according to the HIPAA privacy rules, our results measure the additional information loss resulting solely from masking quasi-identifiers.

In the literature, *information loss* of SDC methods focusing on the *k*-anonymity property is usually measured as the *Sum of Squared Errors* (SSE) [2,5,8,16,27]. It is defined as the sum of squares of the *distances* between the original records and their masked version (Eq. (7)). Hence, the lower the SSE is, the lower the information loss, and the higher the data utility will potentially be. To measure the information loss from a *semantic* perspective, we computed SSE scores using the semantic distance *sd* defined in Eq.1 and SNOMED CT as the knowledge base. Since we are dealing with multivalued data, the average of distances for all attribute values is considered, as follows:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} sd\left(x_{ij}, x_{ij}^{A}\right)}{m} \right)^{2} \tag{7}$$

where n is the number of records in the dataset, each one composed by m attributes, x_{ij} is the original value of the jth attribute of the ith record and x_{ij}^{A} denotes its masked version.

In the following, we evaluate the benefits of our semantic framework regarding the preservation of data semantics in comparison with a non-semantic approach. The three SDC methods introduced in Section 3 will be tested under two different configurations: (1) using classical non-semantic operators and (2) using the semantic operators proposed in our framework. In the first case, values are compared using the equality test (i.e., 0 distance if they are identical and 1 otherwise), whereas the centroid is the most frequent record (i.e., mode). This setting depicts the behaviour of classical anonymising approaches dealing with non-numerical attributes from a non-semantic perspective [9,38]. In the second setting, the three operators proposed in Section 2 are used, configuring a semantically-grounded anonymisation.

Since the three SDC methods aim at fulfilling the *k*-anonymity property, an analysis with respect to the *k* anonymity level has been done. Considering the data distribution shown in Fig. 3, the *k*-level has been set from 2 to 15, so that for $k = 15$, up to a 90% of the total amount of records will be masked.

In order to have a reference point for the analysis of the results of the methods, we have also implemented a naïve algorithm based on *supressing* those records not fulfilling the *k*-anonymity property, that is, records whose value tuples are repeated less than *k* times. As stated in Section 2, even though this approach produces a high information loss, it has been applied in the past to anonymise structured datasets [32,36] and, more specifically, medical data [11,28].

Results of the three SDC algorithms (for semantic and non-semantic settings) and the suppression method are shown in Fig. 4, which denotes the differences in –semantic- SSE scores (Eq. (7)) obtained for the dataset described in Section 4.1 for *k*-values between 2 and 15.

As expected, SSE scores grow as *k*-values increase because, in order to guarantee higher levels of privacy, more records must become indistinguishable and, hence, more changes on the original data are required.

Regarding the non-semantic setting, results show that managing and transforming non-numerical data without considering their semantic features worse preserves the meaning of original data. On the contrary, our semantic approach, that considers both semantics and data distribution, produces significantly lower information loss figures. This shows the benefits of exploiting available medical knowledge bases like SNOMED CT, so that data semantics can be considered (and better preserved) during the anonymisation process.

For some methods the improvement brought by our framework is more significant than for others. The case of microaggregation is the most noticeable, since the semantic framework allows retaining more than a 50% more semantic information than a non-semantic approach. This is coherent, since the microaggregation method heavily relies on semantic operators to group records and to aggregate them. On the other side, the resampling method shows the lowest improvement since it first performs a random sampling of input data, which cannot be optimised from a semantic perspective.

[2] http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctarchive.html.

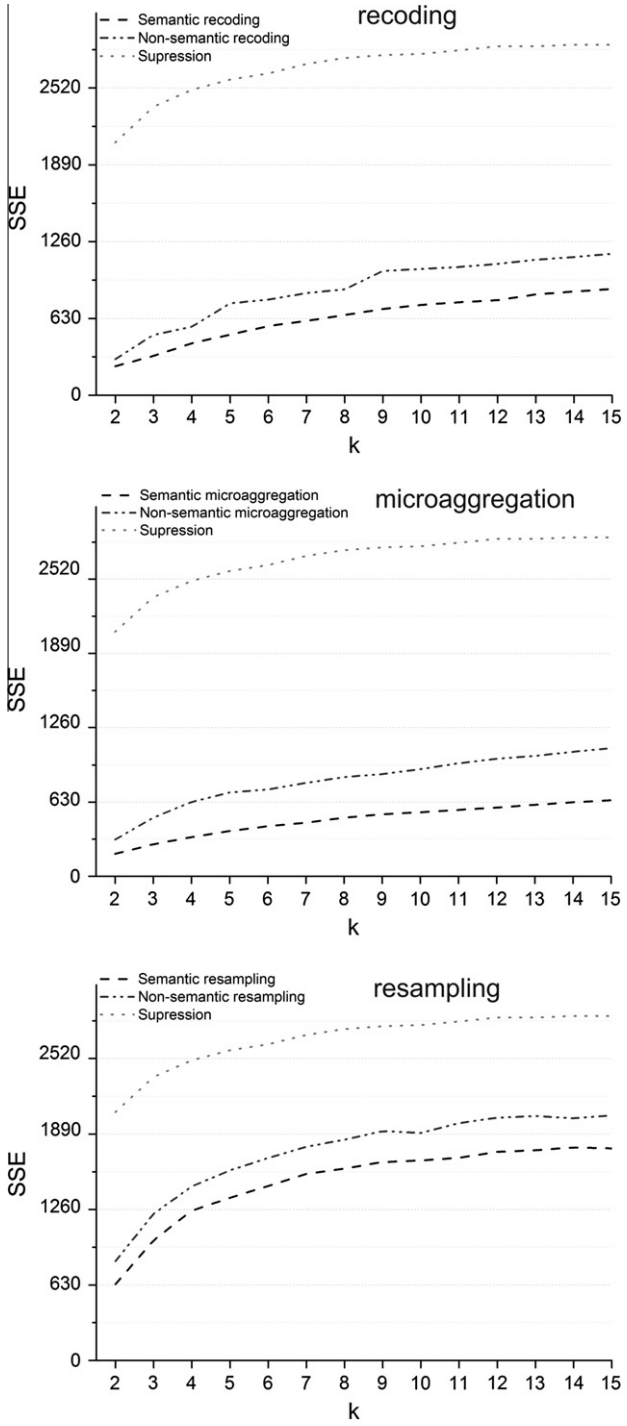

Fig. 4. Semantic Information Loss (SSE) for the three SDC methods (under semantic and non-semantic settings) and the approach based on data suppression for different *k*-anonymity levels.

Information loss obtained when *suppressing* non-anonymous records is significantly higher than what was obtained in the semantic approach. Analysing, for example, SSE values of $k = 15$ (which corresponds to the maximum level of privacy), we can see that for the microaggregation algorithm, the SSE when using the semantic framework is 645, whereas the SSE obtained with suppression is 2875. As discussed in Section 4.1, since most of the records have quite low frequency of appearance in the original EHR, even for low *k*-values, a high percentage of input data is removed. This severely affects data utility, hampering posterior analyses. This shows the importance of applying utility-preserving SDC methods to anonymise the data and, more concretely, taking into consideration both their distributional and semantic features, as the proposed framework pursues.

4.3. Comparing anonymisation methods

This second analysis aims at studying the convenience of using each SDC method for EHR anonymisation when implemented with our semantic framework. The three SDC algorithms have been compared under the perspectives of information loss and runtime.

In addition to the semantic information loss measure introduced above, we also computed a standard utility function focusing solely on the preservation of the original data distribution. In this last case, the well-known *KL-divergence* [14] score has been considered. Being $f(x_i)$ and $f^*(x_i)$ the probability distributions of an original record x_i in the original and masked datasets, respectively, the *KL-divergence* between both datasets is defined as:

$$KL = \sum_{i=1}^{n} f(x_i) \cdot ln \frac{f(x_i)}{f^*(x_i)} \tag{8}$$

A smaller *KL* score indicates a higher similarity between distributions of records between original and masked datasets.

SSE and *KL* scores for the different methods using the proposed framework are shown in Fig. 5 for the same *k* values as above.

In terms of semantic preservation (SSE), the best method is *microaggregation* (which is coherent to results obtained by related works [6,15]) closely followed by *recoding*. This makes sense since both methods heavily rely on semantic operators to aggregate or replace record values. *Resampling*, as discussed above, firstly executes a random data sampling that cannot be optimised from a semantic perspective, hampering data utility. Differences are also reflected on the shape of the SSE function as *k*-values growth. Both *microaggregation* and *recoding* grow almost linearly with respect to *k*. *Resampling*, otherwise, shows an almost logarithmic shape, which is coherent to the fact that the sampling is done on sets of size n/k.

Focusing solely on the preservation of data distribution, *KL* scores show a different picture. In this case, *recoding* provides the best results, followed by *microaggregation* and *resampling*. The fact that both *microaggregation* and *resampling* aggregate records with their centroids, which are synthetically constructed according to the background ontology, may cause that new record values/tuples (i.e. value generalisations and new value tuple combinations) not found in the original dataset appear in the masked version. Even though these new records are semantically similar to original ones, the *KL* score is penalised, since original values are not found in the masked version. This fact significantly alters the probability distribution of masked data with regards to the original one. On the contrary, *recoding* method systematically replaces records for already existing values. Hence, the probability of finding an original record in the masked dataset will increase, resulting in more similar data distributions. *Resampling*, again, provides the worst results (especially for high *k* values) due to the randomness of the sampling process. This produces less cohesive groups and, hence, more general centroids that will more likely correspond to generalisations rather than to values found in the input dataset.

In addition to the degree of information (both semantic and/or distributional) preservation, considering that EHRs are likely to contain large amounts of data, the computational efficiency of data anonymisation is a relevant feature to consider when resources are limited. Fig. 6 shows the comparison for the three SDC methods executed on a 2.4 GHz Intel Core processor with 4 GB RAM.

Runtime figures show that the fastest method is *resampling*, with an almost negligible runtime. *Microaggregation* is the slowest

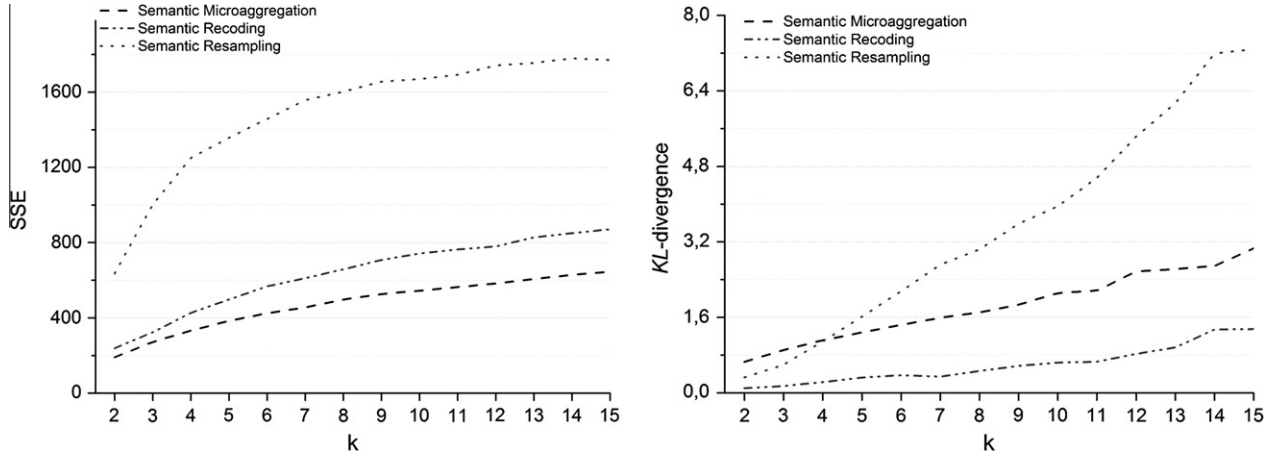

Fig. 5. SSE (semantic) and *KL*-divergence (distributional) scores for the three SDC methods applying the proposed framework across different levels of *k*-anonymity.

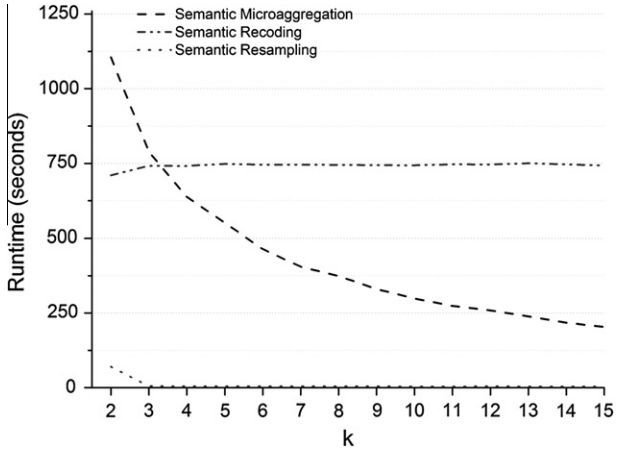

Fig. 6. Runtime (in seconds) for the three SDC methods applying the proposed framework across different levels of *k*-anonymity.

for *k* values below 4, whereas it surpasses *recoding* for higher *k*-values, with an almost inverse logarithmic shape. Results are coherent to the computational costs of the different methods. First, *microaggregation* scales $O(p^2/k)$, where p is the number of distinct records. Since the number of needed clusters lowers as k increases (since each cluster groups, at least, k records), the number of microaggregation iterations and, hence the runtime, lowers as k increases. *Recoding* replaces non-*k*-anonymous records by other similar ones, which implies a cost of $O(p^2)$ for each iteration, resulting in a total of $O(p^3)$ in the worst case (i.e. all records are non-*k*-anonymous). Since the asymptotic computational cost can neglect the value of k, the runtime is almost constant with respect to this parameter.

Finally, as resampling firstly randomly divides input data in n/k samples, it scales linearly as $O(n/k)$, resulting in the lowest runtime, which is even lower when k values increase.

Considering the above results, the semantically-grounded *microaggregation* method seems the best approach to anonymise clinical data when the meaning of original data should be preserved as much as possible. Moreover, it is especially efficient for high *k*-anonymity values. *Recoding* would be only considered if very low *k*-anonymisation levels are required (being more computationally efficient than *microaggregation*) or when analyses to be performed over masked data will be focused solely on data distribution rather than on their semantics. Finally, only when input EHRs are so large to be non-computationally feasibly anonymised by means of *microaggregation* or *recoding* methods, the *resampling* method could be considered thanks to its high efficiency, at the expenses of a higher information loss (both semantic and distributional).

5. Conclusions

Appropriate measures to protect the privacy of EHRs should be taken before making them available for clinical research. Previous works have shown that a basic de-identification of medical data as proposed by the HIPAA Privacy Rule is, sometimes, not enough. On the one hand, original data should be anonymised in a way that the chance of patient disclosure, even when applying statistical methods, is sufficiently reduced. On the other hand, anonymised datasets should retain as much information as possible, so that they are still useful for research tasks. SDC methods aim at balancing these two complementary dimensions, offering additional privacy guarantees with respect to the removal of direct identifiers, which represents their point of departure. However, most of these methods have been designed to deal with numerical values.

Considering the amount and importance of non-numerical data in EHRs, in this paper we presented a general framework that provides semantically-grounded *comparison*, *aggregation* and *sorting* operators. Exploiting a structured medical knowledge base like SNOMED-CT, and relying on the theory of *semantic similarity*, these operators enable a semantically-coherent interpretation of non-numerical attributes, while also considering their distributional features. Since many SDC methods (particularly those focused on fulfilling *k*-anonymity) rely on these basic operators to anonymise data, as described in Section 1.2, they can directly apply the proposed semantic framework to produce semantically-grounded anonymisations for non-numerical data. As shown in the evaluation performed with a real clinical dataset, the use of this framework considerably improves the degree of semantic preservation for all the considered methods, in comparison with non-semantic approaches based solely on information distribution. As a result, the meaning of data and, hence, the utility of the anonymised data from a semantic perspective is better preserved.

As future work, we plan to research on the application of our semantic framework to other privacy models and methods. Particularly it is important to note that privacy models such as *k-anonymity*, even though providing a more robust anonymisation in front of statistical disclosure attacks than the sole removal of identifying attributes, present some limitations. A series of attacks based on the background knowledge that a potential attacker may have about certain individuals have been identified [17]. In these cases, more robust privacy models such as *differential privacy* can be embraced. *Differential privacy* [10] ensures that released

data are insensitive to any individual's data (i.e. the alteration of one input records). Hence, individual data remains uncertain for an attacker. To achieve this, an amount of noise is usually added to the output to introduce uncertainty, a process that implies a loss of information. The approach proposed in this paper could be extended to fulfil differential privacy by adding appropriate noise to, for example, the record counts of each aggregated group [25]. In this case, an equilibrium between the degree of data aggregation and the amount of noise should be carefully considered to minimise information loss.

Another line of research consists on supporting the anonymisation of less structured textual inputs such as free text. In this case, a lexico-syntactic pre-processing stage will be necessary to (1) identify nouns or noun phrases (with semantic content) within the text, and (2) to map them to their corresponding ontological concepts, so that their semantics can be interpreted. Linguistic tools including sentence/token detection, part-of-speech tagging and stemming can be considered.

Acknowledgments

This work was partly funded by the Spanish Government through the projects CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES", eAEGIS TSI2007-65406-C03-02, ICWT TIN2012-32757 and BallotNext IPT-2012-0603-430000 and the Government of Catalonia grant 2009 SGR 1135 and 2009 SGR-01523 and the European Commision under FP7 project Inter-Trust. Sergio Martínez Lluís is supported by a research grant of the Ministerio de Educación y Ciencia (Spain). Evaluation on real patient data has been possible thanks to the *Public-patient Discharge Data* provided by OSHPD, California. The authors acknowledge the collaboration of Xavier Salvadó Martí in the data files pre-processing and in the implementation of part of the software.

References

[1] GPO, US: 45 C.F.R. 164 Security and Privacy 2008. <http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html>.
[2] Abril D, Navarro-Arribas G, Torra V. Towards semantic microaggregation of categorical data for confidential documents. In: Proceedings of the 7th international conference on Modeling decisions for artificial intelligence. Perpignan (France): Springer-Verlag; 2010. p. 266–76.
[3] Batet M, Sanchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 2011;44:118–25.
[4] Chen R, Mohammed N, Fung BCM, Desai BC, Xiong L. Publishing set-valued data via differential privacy. PVLDB 2011;4:1087–98.
[5] Domingo-Ferrer J. A survey of inference control methods for privacy-preserving data mining. In: Aggarwal CC, Yu PS, editors. Privacy-preserving data mining. Springer US; 2008. p. 53–80.
[6] Domingo-Ferrer J, Martínez-Ballesté A, Mateo-Sanz J, Sebé F. Efficient multivariate data-oriented microaggregation. VLDB J 2006;15:355–69.
[7] Domingo-Ferrer J, Mateo-Sanz JM. Resampling for statistical confidentiality in contingency tables. Comput Math Appl 1999;38:13–32.
[8] Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans Knowl and Data Eng 2002; 14:189–201.
[9] Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation. Data Min Knowl Discov 2005; 11:195–212.
[10] Dwork C. Differential privacy. In: ICALP. Springer; 2006. p. 1–12.
[11] Elliot M, Purdam K, Smith D. Statistical disclosure control architectures for patient records in biomedical information systems. J Biomed Inform 2008; 41:58–64.
[12] He Y, Naughton J. Anonymization of set-valued data via top-down, local generalization, VLDB '09: the thirtieth international conference on very large data bases. Lyon, France: VLDB Endowment; 2009.
[13] Jones DH, Adam NR. Disclosure avoidance using the bootstrap and other resampling schemes. In: Proceedings of the fifth annual research conference. Washington (DC): U.S. Bureau of the Census; 1989. p. 446–55.
[14] Kullback S, Leibler R. On information and sufficiency. Ann Math Stat 1951;22:79–86.
[15] Lin J-L, Chang P-C, Liu JY-C, Wen T-H. Comparison of microaggregation approaches on anonymized data quality. Exp Syst Appl 2010;37:8161–5.
[16] Lin J-L, Wen T-H, Hsieh J-C, Chang P-C. Density-based microaggregation for statistical disclosure control. Exp Syst Appl 2010;37:3256–63.
[17] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. *L*-diversity: privacy beyond *k*-anonymity. ACM Trans Knowl Discov Data 2007;1:3.
[18] Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform 2004;37:179–92.
[19] Martínez S, Sánchez D, Valls A. Semantic adaptive microaggregation of categorical microdata. Comput Secur 2012;31:653–72.
[20] Martínez S, Sánchez D, Valls A. Towards *k*-anonymous non-numerical data via semantic resampling. In: Greco S, et al., editors. Information processing and management of uncertainty in knowledge-based systems. Catania, Italy; 2012. p. 519–28.
[21] Martinez S, Sanchez D, Valls A, Batet M. Privacy protection of textual attributes through a semantic-based masking method. Inf Fusion 2011;13:304–14.
[22] Martínez S, Sánchez D, Valls A, Batet M. The role of ontologies in the anonymization of textual variables. In: Proceeding of the 2010 conference on artificial intelligence research and development: proceedings of the 13th international conference of the Catalan association for Artificial intelligence. IOS Press; 2010. p. 153–62.
[23] Martínez S, Valls A, Sánchez D. Semantically-grounded construction of centroids for datasets with textual attributes. Knowl Based Syst 2012;35:160–72.
[24] Meystre S, Friedlin J, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10:70.
[25] Mohammed N, Chen R, Fung BCM, Yu PS. Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. San Diego, California (USA): ACM; 2011. p. 493–501.
[26] Nelson SJ, Johnston D, Humphreys BL. Relationships in medical subject headings. In: Publishers KA, editor. Relationships in the organization of knowledge. New York; 2001. p. 171–84.
[27] Nin J, Herranz J, Torra V. On the disclosure risk of multivariate microaggregation. Knowl Data Eng 2008;67:399–412.
[28] Ohno-Machado L, Silveira PSP, Vinterbo SA. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. Int J Med Inform 2004;73:599–606.
[29] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40:288–99.
[30] Purdam K, Elliot M. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environ Plan A 2007;39:1101–18.
[31] Rogers J. Publically reported breaches in EPR confidentiality; 2005.
[32] Samarati P, Sweeney L. Protecting privacy when disclosing Information: *k*-anonymity and its enforcement through generalization and suppression, technical report SRI-CSL-98-04, SRI Computer Science Laboratory; 1998.
[33] Sanchez D, Batet M. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J Biomed Inform 2011;44:749–59.
[34] Sánchez D, Batet M, Isern D, Valls A. Ontology-based semantic similarity: a new feature-based approach. Exp Syst Appl 2012;39:7718–28.
[35] Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp 1997:640–4.
[36] Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzz Knowl Based Syst 2002;10:571–88.
[37] Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzz Knowl Based Syst 2002;10:557–70.
[38] Torra V. Microaggregation for categorical variables: a median based approach. In: Domingo-Ferrer J, Torra V, editors. Privacy in statistical databases. Berlin (Heidelberg): Springer; 2004. p. 518.
[39] Torra V. Towards knowledge intensive data privacy. In: Proceedings of the 5th international workshop on data privacy management, and 3rd international conference on Autonomous spontaneous security. Athens (Greece): Springer-Verlag; 2011. p. 1–7.